



Quantum-Informed Molecular Learning for Natural Product Chemical Space Through Electronic Structure, Reactivity, Intermolecular Forces, and Data-Efficient Property Prediction

Daniel Fischer^{1*}, Laura Meier¹, Stefan Koch², Thomas Braun³

¹Department of Quantum-Informed Molecular Learning, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany.

²Department of Electronic Structure and Reactivity for Natural Products, Faculty of Pharmacy, Technical University of Munich, Munich, Germany.

³Department of Data-Efficient Property Prediction, Faculty of Pharmacy, University of Kiel, Kiel, Germany.

ABSTRACT

Natural products occupy chemically heterogeneous, stereochemically rich molecular space in which connectivity alone may not encode the electronic determinants of reactivity, conformational response, or intermolecular recognition. Yet adding quantum-mechanical information indiscriminately can increase computational cost without improving prediction. This original computational theory develops an evidence-bounded account of when and how quantum information should enter molecular learning for natural-product chemical space. It distinguishes explicit quantum descriptors, learned electronic representations, electronic intermediates, three-dimensional equivariant representations, and physics-informed interaction models; separates reactivity from mechanistic proof and intermolecular energy components from binding or pharmacological activity; and treats label fidelity and acquisition cost as model-design variables. The central contribution is the proposed Quantum-Informed Natural Product Learning Theory: quantum information should be allocated according to a demonstrable physics deficit in the baseline representation, the electronic sensitivity of the endpoint, the relevant molecular state and environment, and the uncertainty associated with transfer. Under this view, quantum augmentation is neither a universal prerequisite nor an ornamental feature. It is an adaptive escalation strategy that can move from topology and geometry toward electronic structure, polarization, long-range interactions, or higher-fidelity labels only when justified. The theory further links multi-fidelity learning and active acquisition to the scarcity of expensive labels. Its principal boundaries are chemical-space coverage, reference-method error, representation dependence, stereochemistry, charge and protonation state, nonlocality, distribution shift, and uncertainty calibration. Prospective natural-product-specific benchmarks are required before claims of general predictive superiority or translational readiness.

Key Words: Natural products, Quantum machine learning, Electronic structure, Molecular reactivity, Equivariant learning, Multi-fidelity learning

eIJPPR 2026; 16(4):71-79

HOW TO CITE THIS ARTICLE: Fischer D, Meier L, Koch S, Braun T. Quantum-Informed Molecular Learning for Natural Product Chemical Space Through Electronic Structure, Reactivity, Intermolecular Forces, and Data-Efficient Property Prediction. Int J Pharm Phytopharmacol Res. 2026;16(4):71-9. <https://doi.org/10.51847/IJAuys20UL>

INTRODUCTION

Natural-product learning begins with a representation problem before it becomes a quantum problem. Curated natural-product resources assemble structurally diverse molecules whose identity, standardization, provenance,

and annotation quality require explicit control [1]. Even within conventional cheminformatics, different fingerprints organize natural-product chemical space differently, so apparent neighbourhoods, novelty, and similarity are partly representation-dependent rather than intrinsic properties of a database [2]. This representational

Corresponding author: Daniel Fischer

Address: Department of Quantum-Informed Molecular Learning, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany.

E-mail: ✉ daniel.fischer@pharmazie.uni-freiburg.de

Received: 17 April 2026; **Revised:** 25 July 2026; **Accepted:** 27 July 2026

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



instability becomes consequential when model evaluation treats structural proximity as evidence that training and test compounds pose equivalent prediction problems. Natural products commonly combine dense stereochemistry, fused or bridged scaffolds, heteroatom-rich motifs, extended conjugation, macrocyclic flexibility, and uncommon substitution patterns. Those features do not prove that quantum descriptors are necessary. They do, however, make it unsafe to assume that a single topology-centred representation will preserve every distinction relevant to an electronically sensitive endpoint.

The crucial question is therefore not whether quantum mechanics is “more physical” than molecular graphs, but whether an explicit electronic variable contributes information that a baseline representation cannot recover reliably in the available data regime. Direct benchmarking shows that quantum-mechanical descriptors can improve graph-based molecular-property prediction for some tasks, while their value depends on factors including descriptor relevance and dataset size rather than appearing universally [3]. This conditionality is central to the present theory. A costly electronic descriptor that is redundant with molecular topology is not automatically informative; conversely, a modest electronic signal may become valuable when labels are sparse, local chemistry changes sharply across close analogues, or the prediction target depends on charge redistribution, polarization, or bond-making propensity.

A second option is to learn electronic information rather than precompute a fixed catalogue of quantum descriptors. Learned real-space chemical descriptors demonstrate that electronic information can be represented in forms that remain chemically interpretable while complementing structure-centred learning [4]. This supports a broader design space than “calculate descriptors and concatenate them.” The proposed starting principle is consequently selective electronic escalation: first establish molecular identity, stereochemical state, and an appropriate structural baseline; then ask what physical information is missing for the endpoint. Quantum-informed learning is justified only when that missing information is expected to reduce a defined representational deficit and when its added cost can be tested against simpler alternatives. The remainder of the article develops this argument across electronic representations, reactivity, intermolecular forces, equivariant architectures, multi-fidelity learning, expensive-label acquisition, applications, and transfer boundaries.

Quantum descriptors and learned electronic representations

“Quantum-informed” should not be treated as a single representation class. At one end are explicit calculated

quantities attached to atoms, bonds, conformers, or whole molecules. At another are learned electronic objects. Neural models have been constructed to learn wavefunction-derived and orbital quantities, showing that electronic structure itself can function as an intermediate representation rather than only as a source of fixed input descriptors [5]. For natural-product learning, the conceptual importance is not that a learned wavefunction is exact; it is that electronic state can occupy an explicit layer between molecular geometry and a downstream prediction. Its quality necessarily remains conditional on the electronic-structure reference and chemical domain from which that layer was learned.

Electronic information can also enter geometric learning through interactions derived from lower-cost electronic-structure calculations. Orbital-informed geometric deep learning has shown that electronic-interaction features can be embedded into a three-dimensional model to accelerate quantum-chemical prediction within the studied chemical domain [6]. This suggests a hybrid design in which geometry determines where interactions can occur while electronically informed features condition what those interactions mean. Such architectures may be useful when purely structural models conflate geometrically similar environments with electronically different ones, but their natural-product transfer must be demonstrated rather than assumed. An electronic representation therefore inherits approximation error from its reference calculation in addition to statistical error from learning.

A third strategy uses learned electronic intermediates rather than wavefunctions or a conventional descriptor vector. Atom-based electronic populations have been used as intermediate variables in learned potential-energy representations [7]. Together, these approaches motivate a proposed three-part distinction: explicit quantum descriptors, learned electronic fields or wavefunction-related representations, and learned electronic intermediates. The distinction matters because each carries different computational costs, reference-method dependencies, interpretability properties, and transfer risks. It also prevents “quantum-informed” from becoming an imprecise label for any model trained on quantum data. A sufficiently expressive structural model may infer correlated electronic patterns implicitly; explicit electronic representation is warranted when that implicit inference is unreliable, data-inefficient, or scientifically too opaque for the intended task.

Reactivity and bond-making propensity

Reactivity is a particularly strong test of whether electronic information adds something that molecular connectivity alone cannot supply. Quantum-chemistry-augmented neural networks have improved reactivity prediction and

interpretability in studied reaction benchmarks, while generalizability remains dependent on the task and data regime [8]. The relevant lesson is narrower than “quantum descriptors explain reactions.” Reaction labels can contain strong topological and contextual regularities, and a flexible neural model may learn many of them without explicit electronic calculations. Quantum augmentation becomes most defensible when prediction depends on local electron redistribution, competing reactive sites, or chemically close structures that separate sharply in reactivity. Its value should therefore be tested against an equally well-tuned non-quantum baseline.

Regioselectivity illustrates this local-information argument. Learned reaction representations combined with on-the-fly quantum-mechanical descriptors have supported regioselectivity prediction under limited-data conditions [9]. For natural products, the proposed extension is to prioritize site-specific electronic information rather than indiscriminately attaching expensive whole-molecule calculations to every task. A model could therefore represent candidate reactive centres, their local geometric environments, and selected electronic quantities, while preserving reaction conditions and partner identity as separate context. Such a system would predict a labelled transformation or selectivity pattern; it would not establish the molecular mechanism that generated the label.

Three-dimensional reaction representations add a different source of information. Geometric deep learning can encode reaction structure beyond two-dimensional connectivity, but equivariance is not automatically superior on every reaction benchmark [10]. This distinction prevents a common conceptual collapse: geometry, symmetry, and electronic structure solve different representation problems. Stereochemical accessibility or spatial approach may demand three-dimensional encoding; local electrophilicity or nucleophilicity may demand electronic information; neither implies that the other is dispensable. The proposed theory therefore treats reactive prediction as a conditional composition of topology, geometry, electronic state, and reaction context, with each layer justified by ablation against a simpler baseline and evaluated outside the narrow analogue neighbourhoods from which it was learned.

Polarization, dispersion, and intermolecular forces

Intermolecular learning is often compressed into a single target such as interaction energy, but natural-product recognition can involve physically distinct contributions whose importance changes with chemical environment. Neural potentials can jointly predict energies, forces, dipole moments, and partial charges while incorporating physically motivated electrostatic terms [11]. These results establish that electrostatic observables and energetics can

be learned together; they do not make partial-charge assignments unique physical observables, nor do they show that a learned electrostatic term captures every nonlocal response relevant to a complex natural product.

Polarization introduces a different question: how the electronic distribution responds to an external field or nearby partner rather than what static charges are assigned in isolation. High-level quantum reference data have been used to train molecular polarizability models that transfer across molecular sizes within the tested domain [12]. For the present theory, this supports treating inducible response as a separate information requirement. Static topology or fixed-charge features may be adequate for some endpoints, whereas environments involving strong fields, heterogeneous binding sites, or highly polarizable substituents may require a representation capable of changing with context.

Dispersion should likewise remain conceptually distinct. Physics-based functional forms combined with machine learning can model dispersion contributions while retaining an explicit physical decomposition [13]. This does not imply that dispersion determines affinity, selectivity, condensed-phase stability, or biological activity by itself. The proposed computational rule is instead component-aware escalation: use the least complex representation that resolves the physical contribution relevant to the endpoint, and introduce explicit electrostatics, polarization, dispersion, or other nonlocal terms only when simpler models show a defined failure. This preserves interpretability without assuming that every interaction must be decomposed. It also sets a boundary for later applications: improved quantum-quality interaction modeling can strengthen a computational description of molecular recognition, but it must remain separate from binding free energy, target engagement, functional response, and pharmacological effect.

Quantum-informed graph and equivariant models

Three-dimensional molecular learning benefits from separating symmetry from electronic content. E(3)-equivariant message passing can improve data efficiency and accuracy for interatomic-potential learning by encoding how geometric features should transform under rotation, translation, and reflection [14]. This is a structural inductive bias: it reduces the burden of relearning symmetry from examples. For natural products, that can matter when conformationally rich or stereochemically dense structures make orientation-sensitive geometry important. Equivariance, however, does not supply charge redistribution, polarization, or other electronic physics merely by respecting spatial transformations.

Architectures can add those dimensions explicitly. SpookyNet demonstrates that neural force fields can

incorporate electronic degrees of freedom and nonlocal effects rather than leaving all such behavior implicit in local geometry [15]. This motivates a distinction between a model that is geometrically expressive and one that is electronically expressive. A larger receptive field or deeper network may approximate some nonlocal dependencies, but such approximation is not equivalent to representing the relevant physical variable, and its transfer behavior must be tested.

Vector-scalar equivariant message passing provides another route to richer three-dimensional molecular representations for quantum-property prediction [16]. The contribution of such models is best interpreted as structured geometric learning rather than automatic resolution of stereochemistry, protonation, electronic state, or environmental response. Equivariance, electronic/nonlocal modeling and vector-scalar geometric representation address different representation deficits and should not be treated as interchangeable [14–16].

Transfer breadth introduces a further axis. MACE-OFF shows that a high-level quantum-trained, short-range equivariant force field can operate across diverse organic molecular and condensed-phase tasks while retaining an explicitly short-range modeling design [17]. This combination is instructive: broad empirical transfer does not imply physical completeness. The proposed theory therefore separates geometric symmetry, electronic-state information, interaction range, and training-domain coverage. A method should be selected for the deficit it addresses, not because “3D,” “equivariant,” or “quantum” is presumed to dominate simpler representations. In practice, that requires factorial ablations: hold training data and endpoint constant while adding symmetry handling, electronic features, interaction range, or model capacity separately. Only then can an apparent quantum-informed gain be attributed to the layer claimed to supply it rather than to parameter count or training protocol.

These distinctions motivate the method-use-case comparison in **Table 1**.

Table 1. Manuscript-derived comparison of quantum-informed representation strategies and the deficits they are suited to address

Strategy	Best-justified use in this framework	Information added beyond a basic graph	Principal boundary
Explicit QM augmentation	Electronically sensitive property or reactive-site prediction	Calculated local or global electronic descriptors	Redundant cost when topology already captures the signal
Learned electronic intermediate	Quantum-property or potential-energy learning	Wavefunction-, orbital-, or population-related latent information	Reference-method and chemical-domain dependence
3D/equivariant representation	Geometry-sensitive energies, forces, or reactions	Symmetry-consistent spatial relationships	Geometry alone does not encode electronic state
Component/nonlocal model	Electrostatics, polarization, dispersion, or nonlocal force response	Explicit interaction physics or extended interaction range	Approximation and transfer remain component-specific

Note: The use assignments are a proposed synthesis of the manuscript argument; they are not validated universal model-selection rules.

Delta learning and multi-fidelity strategies

Quantum labels are not uniform commodities. Their computational cost and systematic error vary with electronic-structure method, basis, molecular size, state, and conformational sampling. Multi-fidelity learning provides a way to treat these differences as structure in the learning problem rather than discarding cheaper calculations. Molecular excitation-energy studies show that information from multiple quantum-chemical fidelities can be combined to predict higher-fidelity quantities more efficiently than relying only on the most expensive labels [18]. The generalizable premise is not that low fidelity is “almost high fidelity,” but that correlated error can sometimes be learned.

Graph-based transfer learning similarly shows that lower-fidelity labels can improve higher-fidelity molecular-property prediction when the fidelities are sufficiently related [19]. This creates a direct risk of negative transfer: a cheap method that misorders states, mishandles charge,

or fails systematically for a natural-product chemotype can propagate bias rather than reduce variance. The proposed decision is therefore to measure cross-fidelity usefulness before scaling transfer, including ablations in which low-fidelity information is removed.

Curated multi-fidelity resources make accuracy-cost relationships benchmarkable across levels of theory [20]. For natural products, the corresponding design should record not only the nominal method but also molecular state, geometry-generation protocol, convergence status, and the relation between low- and high-fidelity errors. Delta learning is then one special case of a wider principle: learn a correction only when the baseline error has structure that transfers. Where that condition fails, independent high-fidelity sampling or a different representation may be preferable. A natural-product implementation should therefore test whether low- and high-fidelity models preserve the same chemically important ordering across scaffolds, states, and

conformers, and whether the residual error remains learnable after distribution shift. Multi-fidelity learning should consequently be judged by high-fidelity external error, calibration, failure under chemistry shift, and total label-generation cost—not by the volume of inexpensive calculations consumed.

Data-efficient learning with expensive labels

When high-fidelity labels are expensive, the central problem changes from fitting a model to deciding which calculations are worth performing. Physics-informed active learning can prioritize new quantum-chemical calculations so acquisition is more informative than unguided sampling [21]. In natural-product space, this suggests selecting structures, conformers, charge states, or local environments because they expose uncertainty or missing physical regimes rather than because they are abundant. Acquisition remains vulnerable to its own bias: an uncertainty proxy can repeatedly sample familiar boundaries while leaving unseen chemotypes untouched. Small-data molecular-property work further shows that quantum-mechanics-assisted learning can improve extrapolative prediction in a studied limited-data regime [22]. That evidence supports targeted use of expensive

information when empirical labels are scarce, but not the assumption that any quantum descriptor rescues any small dataset. The relevant comparison is against simpler models under the same split, with attention to whether apparent extrapolation is actually supported by close analogues.

Active learning combined with enhanced sampling can also concentrate expensive labels in reactive regions of configuration space [23]. Although that evidence arises from catalytic reactivity rather than biological natural products, the methodological lesson is useful: rare, high-energy, or reactive configurations may require deliberate sampling rather than equilibrium-density coverage. The proposed acquisition rule is therefore to allocate high-cost labels according to three distinct needs—coverage gaps, endpoint-relevant physical uncertainty, and rare reactive/configurational states—while retaining an external holdout that the acquisition loop cannot optimize against. The stopping criterion should likewise be scientific rather than purely numerical: additional calculations are justified when they close a defined coverage or physics gap, not simply because the acquisition score remains nonzero.

Table 2 translates this argument into a trade-off matrix rather than a universal ranking.

Table 2. Proposed decision matrix for allocating quantum-label fidelity and acquisition effort

Situation	Preferred strategy	Expected advantage	Failure condition
Many cheap labels, few expensive labels	Multi-fidelity transfer	Reuse correlated low-fidelity structure	Fidelity mismatch or negative transfer
Structured baseline error	Delta/correction learning	Focus high-fidelity data on residual error	Baseline error changes across chemotypes
Sparse coverage with selectable calculations	Physics-informed active learning	Spend labels on informative gaps	Acquisition proxy misses novel regions
Rare reactive or configurational states	Active learning plus enhanced sampling	Enrich difficult configurations	Sampling regime is not relevant to the target endpoint
Small empirical dataset with plausible electronic signal	Targeted QM-assisted learning	Add physics-linked information	Descriptor is redundant or nontransferable

Note: “Preferred” denotes the proposed choice conditional on the stated situation, not demonstrated superiority across natural-product tasks.

The completed argument through data-efficient learning yields the selective escalation sequence summarized in **Figure 1**.

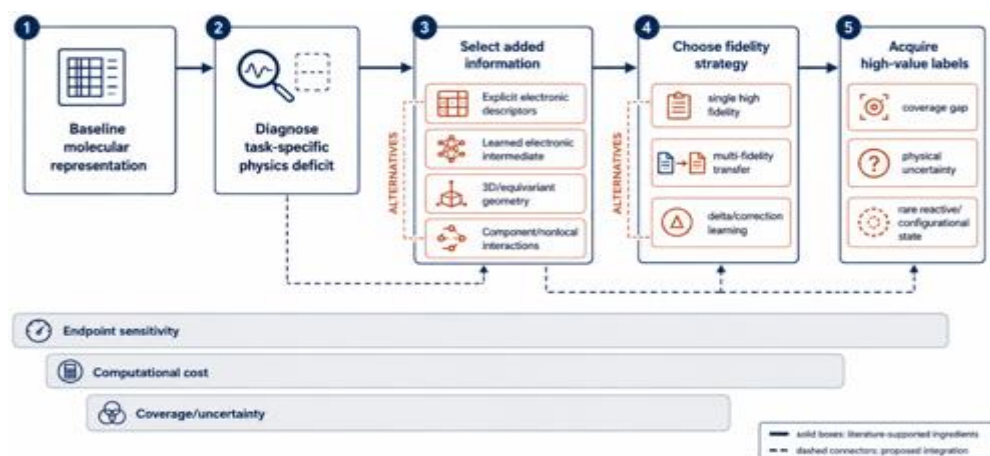


Figure 1. Selective electronic escalation from representation deficit to high-cost label acquisition

Applications across binding, reactivity, and property prediction

Quantum-informed molecular learning becomes useful only when the computational endpoint is defined precisely. In protein–ligand modeling, a physics-aware neural network using electron-density-informed intermediates can approximate quantum-quality interaction energies at substantially lower computational cost than direct quantum treatment [24]. The relevant endpoint is interaction energy. It is not binding free energy, affinity, target engagement, functional activity, or therapeutic efficacy. For natural products, this separation is particularly important because conformational entropy, solvation, protonation, receptor rearrangement, and competing states can intervene between an interaction-energy calculation and an experimental pharmacological outcome.

Reactivity provides a second application class. Quantum-chemistry-derived information has been used to learn experimental solvent effects on reaction rates, demonstrating that reactivity prediction can be conditioned on environment rather than treated as an intrinsic molecule-only label [25]. The proposed implication is that natural-product reactivity models should encode the chemical context relevant to the endpoint—solvent, reaction partner, local environment, or state—rather than interpret a single gas-phase electronic descriptor as universal propensity. Such predictions still do not establish enzymatic selectivity, cellular reactivity, metabolism, or causal biological mechanism.

A third application is atomistic dynamics. Machine-learned quantum force fields trained on chemically diverse fragments can extend quantum-quality dynamics toward larger biomolecular systems [26]. Fragment diversity therefore offers a plausible route to scalable coverage, but transfer to an unseen natural-product scaffold or long-timescale event remains an empirical question. Across all three examples, the article proposes an **endpoint-preservation rule**: the quantum layer should improve the

computational quantity it is trained to approximate, while every higher-level biological or translational inference requires separate evidence. The same discipline applies to ordinary property prediction. A model trained against a computed electronic property should be evaluated as a surrogate for that reference property; a model trained against an experimental endpoint must additionally confront measurement conditions and label heterogeneity. Combining those targets without declaring their evidential level would obscure rather than solve the natural-product prediction problem.

The quantum-informed natural product learning theory

The proposed Quantum-Informed Natural Product Learning Theory begins from molecular-state adequacy. Neural potentials can span neutral and charged organic chemistry, supporting the premise that charge state should be represented rather than silently assumed away [27]. For natural products, the relevant input may likewise depend on protonation, tautomerism, stereochemistry, conformation, or environmental state. The theory therefore asks first whether the molecular state represented to the model corresponds to the state relevant to the target label. Second, it separates local geometric expressivity from long-range electronic response. Polarizable long-range interactions can be integrated explicitly into machine-learning potentials, although current evidence from materials modeling does not validate natural-product prediction directly [28]. The proposed escalation rule is to add long-range or polarization treatment only when the endpoint and failure analysis justify that additional physics, rather than assuming that every molecular task requires the same interaction range.

Third, uncertainty must govern escalation rather than merely decorate predictions. Graph-based molecular design can incorporate uncertainty quantification so predicted value and confidence are not conflated [29]. Charge-state coverage, explicit long-range treatment and

uncertainty estimation are complementary model requirements in the proposed escalation logic [27–29]. The complete theory can therefore be stated as follows: begin with the simplest chemically valid representation; identify a task-specific physics deficit; add the minimum electronic, geometric, interaction, or fidelity layer capable of addressing it; quantify whether the added layer improves external prediction relative to its cost; and escalate calculation or data acquisition when uncertainty or coverage indicates that the current level is insufficient. This is a falsifiable model-selection theory, not an assertion that quantum augmentation will outperform structural learning across natural-product tasks. Its decisive comparison is therefore counterfactual: for the same endpoint and split, does the proposed escalation layer reduce a pre-specified failure of the simpler model, and does that gain persist after the cost of obtaining the added information is counted?

Computational cost and transfer boundaries

Quantum-informed models are attractive partly because they can compress expensive electronic-structure information into cheaper inference. Orbital-informed learning has approached DFT-level accuracy at semi-empirical-like cost on biological and organic chemistry benchmarks [30]. Such results establish a cost–accuracy opportunity within tested domains, not an entitlement to DFT-quality predictions for arbitrary natural products. The full computational ledger should include descriptor generation, conformer enumeration, state preparation, failed calculations, training, inference, and any high-fidelity labels used for calibration.

Transfer is the harder boundary. Stress testing of machine-learning force fields across heterogeneous systems shows that performance is distribution- and task-dependent, undermining conclusions based only on average in-domain accuracy [31]. Natural-product benchmarks should therefore separate random interpolation from scaffold-, family-, stereochemistry-, charge-state-, and conformer-held-out evaluation. A model that performs strongly on close analogues but fails under those shifts may still be useful, but its claim is local interpolation rather than general chemical transfer.

Long-range physics also creates a cost decision. Charge equilibration can couple local equivariant representations to nonlocal molecular interactions, providing one structured compromise between local efficiency and extended electronic response [32]. It does not capture every many-body polarization or dispersion effect. The proposed evaluation principle is thus two-dimensional: report predictive error together with the incremental computational cost of the physical layer producing it. External natural-product panels should then test whether

the added expense buys robust transfer under the chemical states and environments that matter for the intended endpoint. Cost reporting should be normalized to the decision: library screening, short-list refinement, reactive-site estimation, and dynamics tolerate different per-molecule budgets. Consequently, the relevant frontier is not a single accuracy score but the best attainable error at a stated computational budget and transfer regime. A more expensive quantum layer is justified only when it changes that frontier, not when it merely improves an in-domain benchmark by an amount disconnected from the intended use.

Limitations

The theory depends on knowing when available data are insufficient. Model-free information-theoretic diagnostics can estimate dataset completeness, outliers, and uncertainty-relevant coverage before a fitted predictor is trusted [33], but completeness in one representation does not guarantee completeness for a downstream task. Dataset support and model uncertainty must therefore remain distinct.

Interpretability analyses of reaction models show that dataset and representation biases can masquerade as learned chemistry [34]. Quantum-derived inputs do not remove this problem: biased labels, duplicated analogues, inappropriate splits, or correlated computational artefacts can simply be learned with greater physical-looking sophistication.

Natural-product stereochemistry supplies a specific boundary. Machine-learning stereochemical assignment is sensitive to representation and training-set coverage [35]. Dataset coverage, learned bias and stereochemical representation dependence are distinct transfer risks that require separate tests [33–35]. The present theory therefore requires stereochemistry-, analogue-, scaffold-, and state-aware external evaluation where relevant. These three risks are not exhaustive; conformational sampling, reference-method error, protonation/tautomer ambiguity, environmental mismatch, uncertainty miscalibration, and experimental-label noise can also limit inference. The framework remains a proposed computational theory until prospective natural-product-specific comparisons show when selective quantum escalation succeeds, fails, or offers no advantage over simpler learning. Such tests should predefine failure categories, preserve untouched external families, and report cases in which the simpler baseline is equal or better. Negative results are essential because the theory predicts that some endpoints should not warrant quantum escalation.

CONCLUSION

Quantum-informed learning for natural products should not be framed as a contest between molecular graphs and quantum chemistry. Its more useful question is conditional: what physical information is missing from the simplest chemically valid representation for a defined endpoint, and is that information worth its computational and data cost? This article proposes selective electronic escalation as the organizing answer. Explicit descriptors, learned electronic intermediates, equivariant geometry, polarization or nonlocal interactions, multi-fidelity transfer, and active label acquisition become alternative interventions for different representational deficits rather than mandatory components of one maximal model.

The theory preserves boundaries that are essential for scientific interpretation. Reactivity prediction is not mechanistic proof; interaction energy is not binding affinity or pharmacological activity; quantum-reference agreement is not experimental validity; and uncertainty is not evidence of truth. Natural-product stereochemistry, charge and protonation state, conformational coverage, chemical-family shift, environmental context, reference-method error, and label bias all constrain transfer. The proposed framework is therefore valuable only if it can be challenged. Its appropriate next test is prospective comparison against simpler structural baselines on externally separated natural-product chemical space, with fidelity, uncertainty, computational cost, and endpoint-specific experimental validation reported together. Success would be conditional evidence for selective escalation, not validation of quantum augmentation as a universal rule.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Res.* 2025;53(D1):D634-D43. doi:10.1093/nar/gkae1063
- [2] Boldini D, Ballabio D, Consonni V, Todeschini R, Grisoni F, Sieber SA. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *J Cheminform.* 2024;16(1):35. doi:10.1186/s13321-024-00830-3
- [3] Li SC, Wu H, Menon A, Spiekermann KA, Li YP, Green WH. When do quantum mechanical descriptors help graph neural networks predict chemical properties? *J Am Chem Soc.* 2024;146(33):23103-20. doi:10.1021/jacs.4c04670
- [4] Gallegos M, Vassilev-Galindo V, Poltavsky I, Martín Pendás Á, Tkatchenko A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nat Commun.* 2024;15(1):4345. doi:10.1038/s41467-024-48567-9
- [5] Schütt KT, Gastegger M, Tkatchenko A, Müller KR, Maurer RJ. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat Commun.* 2019;10(1):5024. doi:10.1038/s41467-019-12875-2
- [6] Qiao Z, Christensen AS, Welborn M, Manby FR, Anandkumar A, Miller TF 3rd. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proc Natl Acad Sci U S A.* 2022;119(31):e2205221119. doi:10.1073/pnas.2205221119
- [7] Xie X, Persson KA, Small DW. Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations. *J Chem Theory Comput.* 2020;16(7):4256-70. doi:10.1021/acs.jctc.0c00217
- [8] Stuyver T, Coley CW. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *J Chem Phys.* 2022;156(8):084104. doi:10.1063/5.0079574
- [9] Guan Y, Coley CW, Wu H, Ranasinghe D, Heid E, Struble TJ, et al. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem Sci.* 2021;12(6):2198-208. doi:10.1039/D0SC04823B
- [10] van Gerwen P, Briling KR, Bunne C, Somnath VR, Laplaza R, Krause A, et al. 3DReact: Geometric deep learning for chemical reactions. *J Chem Inf Model.* 2024;64(15):5771-85. doi:10.1021/acs.jcim.4c00104
- [11] Unke OT, Meuwly M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J Chem Theory Comput.* 2019;15(6):3678-93. doi:10.1021/acs.jctc.9b00181
- [12] Wilkins DM, Grisafi A, Yang Y, Lao KU, DiStasio RA Jr, Ceriotti M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc Natl Acad Sci U S A.* 2019;116(9):3401-6. doi:10.1073/pnas.1816132116
- [13] Villot C, Lao KU. Ab initio dispersion potentials based on physics-based functional forms with machine learning. *J Chem Phys.* 2024;160(18):184103. doi:10.1063/5.0204064



- [14] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun.* 2022;13(1):2453. doi:10.1038/s41467-022-29939-5
- [15] Unke OT, Chmiela S, Gastegger M, Schütt KT, Sauceda HE, Müller KR. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat Commun.* 2021;12(1):7273. doi:10.1038/s41467-021-27504-0
- [16] Wang Y, Wang T, Li S, He X, Li M, Wang Z, et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nat Commun.* 2024;15(1):313. doi:10.1038/s41467-023-43720-2
- [17] Kovács DP, Moore JH, Browning NJ, Batatia I, Horton JT, Pu Y, et al. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. *J Am Chem Soc.* 2025;147(21):17598-611. doi:10.1021/jacs.4c07099
- [18] Vinod V, Maity S, Zaspel P, Kleinekathöfer U. Multifidelity machine learning for molecular excitation energies. *J Chem Theory Comput.* 2023;19(21):7658-70. doi:10.1021/acs.jctc.3c00882
- [19] Buterez D, Janet JP, Kiddle SJ, Oglic D, Lió P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat Commun.* 2024;15(1):1517. doi:10.1038/s41467-024-45566-8
- [20] Vinod V, Zaspel P. QeMFi: A multifidelity dataset of quantum chemical properties of diverse molecules. *Sci Data.* 2025;12(1):202. doi:10.1038/s41597-024-04247-3
- [21] Hou YF, Zhang L, Zhang Q, Ge F, Dral PO. Physics-informed active learning for accelerating quantum chemical simulations. *J Chem Theory Comput.* 2024;20(18):7744-54. doi:10.1021/acs.jctc.4c00821
- [22] Shimakawa H, Kumada A, Sato M. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. *NPJ Comput Mater.* 2024;10(1):11. doi:10.1038/s41524-023-01194-2
- [23] Perego S, Bonati L. Data efficient machine learning potentials for modeling catalytic reactivity via active learning and enhanced sampling. *NPJ Comput Mater.* 2024;10(1):291. doi:10.1038/s41524-024-01481-6
- [24] Glick ZL, Metcalf DP, Glick CS, Spronk SA, Koutsoukas A, Cheney DL, et al. A physics-aware neural network for protein-ligand interactions with quantum chemical accuracy. *Chem Sci.* 2024;15(33):13313-24. doi:10.1039/D4SC01029A
- [25] Chung Y, Green WH. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chem Sci.* 2024;15(7):2410-24. doi:10.1039/D3SC05353A
- [26] Unke OT, Stöhr M, Ganscha S, Unterthiner T, Maennel H, Kashubin S, et al. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Sci Adv.* 2024;10(14):eadn4397. doi:10.1126/sciadv.adn4397
- [27] Anstine DM, Zubatyuk R, Isayev O. AIMNet2: A neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem Sci.* 2025;16(23):10228-44. doi:10.1039/D4SC08572H
- [28] Gao R, Yam CY, Mao J, Chen S, Chen GH, Hu Z. A foundation machine learning potential with polarizable long-range interactions for materials modelling. *Nat Commun.* 2025;16(1):10484. doi:10.1038/s41467-025-65496-3
- [29] Chen LY, Li YP. Uncertainty quantification with graph neural networks for efficient molecular design. *Nat Commun.* 2025;16(1):3262. doi:10.1038/s41467-025-58503-0
- [30] Christensen AS, Sirumalla SK, Qiao Z, O'Connor MB, Smith DGA, Ding F, et al. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J Chem Phys.* 2021;155(20):204103. doi:10.1063/5.0061990
- [31] Poltavsky I, Charkin-Gorbunin A, Puleva M, Fonseca GC, Batatia I, Browning NJ, et al. Crash testing machine learning force fields for molecules, materials, and interfaces: Model analysis in the TEA Challenge 2023. *Chem Sci.* 2025;16(8):3720-37. doi:10.1039/D4SC06529H
- [32] Fuchs P, Sanocki M, Zavadlav J. Learning non-local molecular interactions via equivariant local representations and charge equilibration. *NPJ Comput Mater.* 2025;11(1):287. doi:10.1038/s41524-025-01790-4
- [33] Schwalbe-Koda D, Hamel S, Sadigh B, Zhou F, Lordi V. Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning using information theory. *Nat Commun.* 2025;16(1):4014. doi:10.1038/s41467-025-59232-0
- [34] Kovács DP, McCorkindale W, Lee AA. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat Commun.* 2021;12(1):1695. doi:10.1038/s41467-021-21895-w
- [35] Orsi M, Reymond JL. Assigning the stereochemistry of natural products by machine learning. *J Cheminform.* 2026;18(1):76. doi:10.1186/s13321-026-01205-6

